PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-007447
(43)Date of publication of application : 12.01.1999

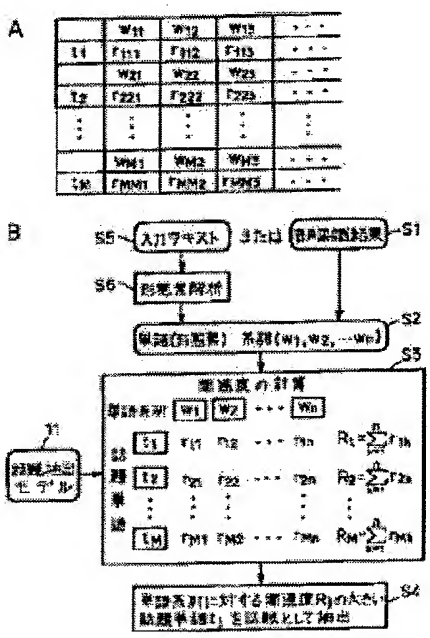
(51)Int.Cl. G06F 17/27
G10L 3/00

(21)Application number : 09-160954
(22)Date of filing : 18.06.1997

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>
(72)Inventor : OFU KATSUTOSHI
MATSUOKA TATSUO
MATSUNAGA SHOICHI

(54) TOPIC EXTRACTING METHOD, TOPIC EXTRACTION MODEL TO BE USED FOR THE EXTRACTING METHOD, PREPARING METHOD FOR THE TOPIC EXTRACTION MODEL, AND TOPIC EXTRACTION PROGRAM RECORDING MEDIUM

(57)Abstract:
PROBLEM TO BE SOLVED: To appropriately extract a topic (heading) expressing the contents of a continuous voice.
SOLUTION: A continuous large vocabulary voice is recognized (SI) and a word sea prepared (S2) by using a model 11 stored by executing syntax analysis for the headings and texts of many newspaper topics, obtaining respective topic words of the headings and words in the texts, finding out the appearance frequency of respective words and the cooccurrence frequency of combination of each topic word and a text word in the same news item, and finding out the degree of association between the topic word and the text word by mutual information volume or an x2 method. The degree of association between each topic word and each word in the word sequence is found out by the model 11 to prepare a association degree sequence, the sum of association degrees in respective association degree sequences is found out (S3) and a topic word corresponding to the maximum sum is outputted (S4).



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平11－7447

(43)公開日　平成11年(1999)１月12日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/27		Ｇ０６Ｆ　15/38	Ｎ
Ｇ１０Ｌ　3/00	５６１	Ｇ１０Ｌ　3/00	５６１Ｇ

審査請求　未請求　請求項の数12　ＯＬ　（全　６　頁）

(21)出願番号　特願平9－160954

(22)出願日　平成９年(1997)６月18日

(71)出願人　000004226
日本電信電話株式会社
東京都新宿区西新宿三丁目19番２号

(72)発明者　大附　克年
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　松岡　達雄
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(72)発明者　松永　昭一
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内

(74)代理人　弁理士　草野　卓

(54)【発明の名称】　話題抽出方法及びこれに用いる話題抽出モデルとその作成方法、話題抽出プログラム記録媒体

(57)【要約】

【課題】　連続音声の内容を表わす話題（見出し）を適切に抽出する。

【解決手段】　大量の新聞記事の見出しと本文とを形態素解析し、その見出しの各話題単語と文中単語を得、その各出現頻度と、同一記事で話題単語と文中単語の組合せの共起頻度を求め、相互情報量又はx^2法により話題単語と文中単語との関連度を求めて格納したモデル１１を用い、連続大語彙音声を音声認識し（Ｓ１）、単語系列を作り（Ｓ２）、各話題単語と単語系列の各単語との関連度をモデル１１が求めて、関連度系列を作り各関連度系列における関連度の和を求め（Ｓ３）、その和の最大のものと対応する話題単語を出力する（Ｓ４）。

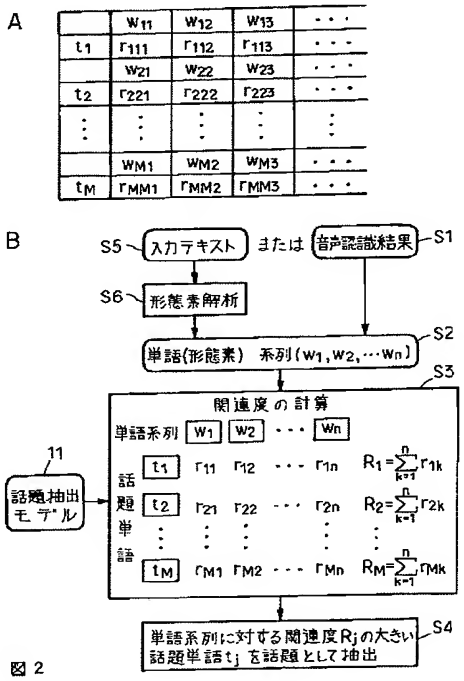


図２

1

【特許請求の範囲】

【請求項１】　複数の単語の系列の内容を表す話題単語を抽出するために用いられるモデルであって、
複数の話題単語と、その各話題単語と、各単語との関連度とがそれぞれ格納されている話題抽出モデル。

【請求項２】　話題単語と各単語との関連度は、話題単語と複数単語との関連度であることを特徴とする請求項１記載の話題抽出モデル。

【請求項３】　上記関連度は話題単語と、各単語との相互情報量に基づくものであることを特徴とする請求項１又は２記載の話題抽出モデル。

【請求項４】　上記関連度は話題単語と各単語とのx^2ベクトル法にもとづくものであることを特徴とする請求項１記載の話題抽出モデル。

【請求項５】　請求項１乃至４の何れかに記載した話題抽出モデルを用いて入力された複数の単語の系列の内容を表す話題単語を抽出する方法であって、
上記話題抽出モデル中の各話題単語ごとに、これと上記入力単語系列の各単語との関連度を上記話題抽出モデルを参照して求めて関連度系列をそれぞれ作り、
これら各関連度系列の各関連度の和を求めて上記単語系列に対する各話題単語の関連度を求め、
これら単語系列に対する関連度中の大きいものから順にＱ個（Ｑは１以上の整数）のものとそれぞれ対応する話題単語を出力することを特徴とする話題抽出方法。

【請求項６】　上記関連度系列の各関連度に対し、これと対応する単語の尤度で重み付けて上記各関連度の和を求めることを特徴とする請求項５記載の話題抽出方法。

【請求項７】　連続した音声信号を単語音声認識して、上記入力単語系列を得ることを特徴とする請求項５又は６記載の話題抽出方法。

【請求項８】　上記認識結果として複数の上位の候補系列を上記入力単語系列とすることを特徴とする請求項７記載の話題抽出方法。

【請求項９】　入力テキストを、形態素解析し、その解析結果の形態素を上記入力単語系列とすることを特徴とする請求項５又は６に記載の話題抽出方法。

【請求項１０】　本文とその見出しよりなる多数のテキストを学習データとし、この学習データの見出し、本文をそれぞれ形態素解析して、見出しの形態素としての話題単語と、本文の形態素としての文中単語を得る工程と、
上記各話題単語の出現頻度と、上記各文中単語の出現頻度と、１つのテキスト中の上記話題単語と上記文中単語の各組み合せが同時に得られる共起頻度とをそれぞれ計数する工程と、
上記話題単語の出現頻度と文中単語の出現頻度と各共起頻度とを用いて各話題単語と各文中単語との関連度を求めて話題抽出モデルを得る工程とを有する話題抽出モデル作成方法。

2

【請求項１１】　上記出現単語中の出現頻度が所定値以下のものを省略し、上記文中単語中の出現頻度の順位が所定値以下のものを省略し、上記出現単語及び上記文中単語中の情報検索という観点から意味的情報を伝達する名詞・動詞などの品詞のもの以外を省略し、かつ１つのテキスト中に出現する上記話題単語及び文中単語の組み合せが所定回数以下の組み合せを省略して残りの話題単語及び文中単語を用いて上記関連度を求めることを特徴とする請求項１０記載の話題抽出モデル作成方法。

【請求項１２】　入力音声を連続音声認識して入力単語系列を得、
複数の話題単語と、その各話題単語と、各単語との関連度とがそれぞれ格納された話題抽出モデルを参照して、上記話題単語ごとにこれと上記入力単語系列中の各単語との関連度を求めて関連度系列を得、
上記各関連度系列から、上記各話題単語ごとの上記入力単語系列の関連度を求め、
これら入力単語系列の関連度中の関連度が最大のものから順にＱ個（Ｑは１以上の整数）のものとそれぞれ対応する話題単語を上記入力単語系列の内容を表わす話題として出力することをコンピュータを用いて行うためのプログラムを記録した記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】この発明は、連続発声された音声の単語認識結果やテキストを形態素解析により分割された単語などの単語列に対し、その内容を表わす話題を抽出する方法、その話題抽出に用いる話題抽出モデルとそのモデルの作成方法に関する。

【０００２】

【従来の技術】連続発声された音声からのその内容を表わす話題抽出では５～１０種類の分野のうちのいずれかの分野に依存度の高いキーワードを予め選択しておき、それらのキーワードを音声区間中から検出（キーワードスポッティング）して、検出されたキーワード集合が最も高い依存度を示す話題を結果として出力する方法が多くとられる。例えば横井、河原、堂下、“キーワードスポッティングに基づくニュース音声の話題同定”、情処研報、SLP6－３、1995. 櫻井、有木、“キーワードスポッティングによるニュース音声の索引付けと分類”、信学技法、SP96－66、1996. R.C.Rose, E.L.Chang , and R.P.Lippmann, “Techniques for Information Retrieval from Voice Messages”, Proc.ICASSP-91, pp.317-320, 1991. などに示されている。

【０００３】また従来の文章（テキスト）から話題を抽出する方法は文中の特定の個所を抽出して行うもので、その処理が複雑であった。

【０００４】

【発明が解決しようとする課題】従来の連続音声の話題抽出方法では、限られた数のキーワードしか用いること

ができず、またキーワードの数を増やした場合には、誤って検出されるキーワードが増えてしまう、また話題の分野が少ないため、情報検索や索引付けに使うことができないという問題があった。また従来のテキストよりの話題抽出方法は、特定の個所を探して行うため、その処理が複雑であった。これを連続音声の話題抽出に適用すると、その所定個所についての単語認識を誤ると、話題抽出は誤ってしまう。

【０００５】この発明の目的は比較的簡単な処理で話題を抽出することができる話題抽出方法、そのプログラムを記録した記録媒体と、上記話題抽出に用いる話題抽出モデルと、その作成方法を提供することにある。

【０００６】

【課題を解決するための手段】この発明の話題抽出モデルは、本文とその見出しよりなるテキストを多数用いて、それぞれ形態素解析を用い、本文単語と話題単語（見出し中の）と得、これら本文単語の出現頻度、話題単語の出現頻度、１テキスト中に本文単語と話題単語の組み合せが同時に存在する共起頻度をそれぞれ求め、これら頻度から各話題単語と、本文単語との関連度を求め、これらを話題抽出モデルとして格納しておく。

【０００７】この発明の話題抽出方法では前記この発明の話題抽出モデルを用い、入力音声の音声認識や入力テキストの形態素解析で、入力単語系列を得、各話題単語と入力単語系列中の各単語との関連度とを話題抽出モデルを参照して求めて話題単語ごとの関連度系列を得、これら関連度系列から各話題単語の入力単語系列との関連度をそれぞれ求め、これら入力単語系列の関連度中の大きいものと対応する話題単語を入力音声又はテキストに対する話題として出力する。

【０００８】この発明の記録媒体はこの発明の話題抽出方法をコンピュータで実行させるためのプログラムが記録されている。

【０００９】

【発明の実施の形態】まずこの発明の話題抽出モデルと＊この作成方法の実施例を説明する。話題抽出モデルの学習（作成）はある話題について述べられているテキストとその内容を表わす複数の話題単語との組を大量に用いて行う。一例として新聞記事の本文と見出しを用いて話題抽出モデルを学習（作成）する場合、約５年分の新聞記事よりその見出しと本文とをそれぞれ取出し（Ｓ１）、これらを形態素解析を行い（Ｓ２）、単語（形態素）に分割し、見出しの形態素（話題単語）と、本文の形態素（文中単語）とを得る。

【００１０】これら話題単語と文中単語について、大量のデータにおける出現頻度と、共起頻度とを用いて、文中単語と話題単語との関連度を求める。しかし、文中単語と話題単語の組み合わせは非常に莫大な数になる。従ってこの実施例では話題単語については、出現回数が２回以上の単語に限り（Ｓ３）、文中単語については出現頻度が上位１５万の単語のみを選出し（Ｓ４）、更に情報検索という観点からより意味的情報を伝達すると考えられる名詞、動詞などの内容語に着目し、ここでは話題単語、文中単語の何れについても名詞、動詞、形容詞、形容動詞、副詞のみを取出す（Ｓ５）。更に話題単語と文中単語との組合せで同一記事に出現するのが１度しかなかったものは除外し、つまり話題単語と文中単語の組み合わせで同一記事に出現することが２回以上のもののみを残した（Ｓ６）。このようにして話題単語の総頻度１２．３×10^6が６．３×10^6となり総数１３６×10^3が７４×10^3となり、文中単語の総頻度２１８．８×10^6が９０．１×10^6となり総数６４０×10^3が１４７×10^3となり、２回以上起きた共起の組み合わせは約５８００万種類となった。

【００１１】この約５８００万種類について、これら単語の出現頻度と共起頻度を用いて文中単語と話題単語との関連度を求める。文中単語w_iと話題単語t_jとの関連度は以下のようにして求める。相互情報量に基づく関連度

$$I(w_i : t_j) = \log(P(w_i, t_j) / P(w_i) P(t_j)) \quad \cdots(1)$$

$P(w_i, t_j)$：w_iとt_jが同時に出現する確率
$P(w_i)$：w_iの出現確率、$P(t_i)$：t_jの出現確率

x^2法に基づく関連度

$x_{ij}^2 = (f_{ij} - F_{ij})^2 / F_{ij}$

【００１２】

【数１】

$$F_{ij} = \frac{\sum_{\ell=1}^{M} f_{i\ell}}{\sum_{k=1}^{N}\sum_{\ell=1}^{M} f_{k\ell}} \cdot \sum_{k=1}^{N} f_{kj}$$

※Ｎ：文中単語の種類数、Ｍ：話題単語の種類数、
f_{ij}：話題単語t_jに対する文中単語w_iの頻度、
F_{ij}：話題単語t_jに対する文中単語w_iの理論（期待）度数

相互情報量の計量において、学習データ中に文中単語w_iと話題単語t_iの共起が観測されない場合、$P(w_i, t_i) = 0$となり、関連度の合計を求める際に問題が生じる。そこで、共起が観測されなかった場合には情報が得られなかったものとして、実際には次式のように相互情報量に基づく関連度を計算する。

【００１３】

※

$I'(w_i : t_j) = I(w_i : t_j)$，$P(w_i, t_j) \neq 0$の場合

0，

一方、x^2 法における理論度数F_{ij}とは、全ての話題単語に対して文中単語w_iが等確率で出現した場合の出現頻度である。実際の出現頻度と理論度数とのずれが大きければ、その文中単語はその話題単語に対して偏って出現していることになる。しかし、上述のx^2 法の式では、実際の出現頻度f_{ij}が理論度数F_{ij}より小さい場合にも、関連度が正の値となってしまうため、実際には次式のようにx^2 法に基づく関連度を計算する。

【００１４】

$$x_{ij}{}^{2'} = x_{ij}{}^2 \quad , \quad f_{ij} - F_{ij} \geq 0 \text{の場合}$$
$$0 \quad , \quad f_{ij} - F_{ij} < 0 \text{の場合}$$

従って、ステップＳ６で得られた文中単語w_iと話題単語t_jとの各組み合せについて、その各頻度Ｐ（w_i）：Ｐ（t_j），Ｐ（w_i，t_j）、またはf_{ij}，をそれぞれ演算し（Ｓ７）、頻度テーブル１１に格納する。これを学習データが終るまで行う（Ｓ８）。学習データが終ると、頻度テーブル１１内に演算した頻度を用いて関連度Ｉ（w_i，t_j）又はF_{ij}の計算を行って話題抽出モデルを得る（S_9）。

【００１５】従って話題抽出モデルは図２Ａに示すように、話題単語の種類t_1，t_2，…t_Nそれぞれについて、これと２回以上共起する文中単語、つまりt_1についてはw_{11}，w_{12}，w_{13}，…との関連度r_{111}，r_{112}，r_{113}，…が、またt_2についてはw_{21}，w_{22}，w_{23}，…との関連度r_{211}，r_{212}，r_{213}，…が、以下同様に文中単語との関連度が格納されている。

【００１６】次にこの話題抽出モデルを用いて連続入力単語列から話題を抽出する方法を図２Ｂを参照して説明する。連続発声される音声を入力とする場合、その入力音声を単語音声認識し（Ｓ１）、認識結果として単語系列w_1，w_2，…w_nを得る（Ｓ２）、これら単語系列w_1，w_2，…w_nの各単語について、話題抽出モデル＊

$$I(x_1:x_2:\cdots:x_n) = \log \frac{[\Pi P(x_i, x_j)] \cdot [\Pi P(x_j, x_j, x_k, x_t)]}{[\Pi P(x_i)] \cdot [\Pi P(x_i, x_j, x_k)]} \quad \cdots (2)$$

Πは、あい異なる添字の全ての組み合せについて計算する。従ってx_1，x_2，…，x_n中１つの話題単語と他のｎ－１個を文中単語との相互情報量をＩ（x_1：x_2：…：x_n）により求めることができる。このように、複数の文中単語と１つの話題単語との関連度を求めておくと、例えば「コンピュータ」と「インターネット」の関連度、また「ネットワーク」と「インターネット」の関連度はそれ程大きくないが、「コンピュータ」と「ネットワーク」が同じ文中にあった場合の「インターネット」への関連度が大きくなるような話題抽出モデルの学習ができる。つまり式（１）の関連度では話題と

Ｐ（w_i，t_j）＝０の場合

＊１１を参照して、その各話題単語t_1，t_2，…t_Nに対する関連度を求める。つまり認識単語w_1に対する話題単語t_1，t_2，…t_Nとの各関連度r_{11}，r_{21}，…，r_{N1}を求め、単語w_2に対する話題単語t_1，t_2，…t_Nとの各関連度r_{12}，r_{22}，…，r_{N2}を求め、以下同様に求める。

【００１７】各話題単語t_1，t_2，…，t_Nについての各認識単語w_1，w_2，…，w_nとの関連度の合計、つまり単語系列に対する関連度R_jを計算する。即ち、話題単語t_1についてはr_{11}，r_{12}，…，r_{1n}の和$R_1 = \Sigma_{k=1}^{n} r_{1k}$を求め、$t_2$については$r_{21}$，$r_{22}$，…，$r_{2n}$の和$R_2 = \Sigma_{k=1}^{n} r_{2k}$を求め、以下同様に$R_3$，…，$R_N$を求める（Ｓ３）。これら単語系列に対する関連度$R_1$，…，$R_N$中で関連度が大きいものから順にＱ個（Ｑは１以上の整数）のものとそれぞれ対応する話題単語t_jの具合を、その単語系列に対する話題とする（Ｓ４）。Ｑは１でもよいが、通常は複数で例えば５程度である。関連度R_1，…，R_N中の大きいものから順に対応する話題単語の複数個を候補として出力してもよい。

【００１８】単語系列から話題の抽出としてはテキストを入力し（Ｓ５）、これを形態素解析し（Ｓ６）、形態素つまり単語列w_1，w_2，…，w_nを得て、これを音声入力の場合と同様に話題抽出モデル１１を用いて処理して、テキストに対する話題を抽出することもできる。関連度をw_iとt_jの相互情報量に基づいて求める場合は式（１）、つまり２点間の相互情報量に基づいて決めた。一方、ｎ点間の相互情報量は次式で定義される。

【００１９】

【数２】

して「インターネット」を抽出できない場合に、式（２）の関連度によると「インターネット」を話題として抽出でき、適切な話題を抽出することができることがある。

【００２０】話題単語t_kと単語系列w_1，w_2，…，w_nとの関連度R_kはt_kに対する各単語の関連度の和$r_{k1} + r_{k2} +$，…，$+ r_{kn}$で求められるが、その加算の際に各単語に対する重みs_1，s_2，…，s_nをそれぞれ付けて、$r_{k1} \times s_1 + r_{k2} \times s_2 +$，…，$+ r_{kn} \times s_n$というようにして、より適切な関連度$R_k$を得るようにすることもできる。ここで重み$s_1$，$s_2$，…，$s_n$

としては、各単語w_1，w_2，…，w_nのその音声認識時の単語の確からしさ（音響的尤度）や言語的尤度、つまりその単語がその前の単語に対し、文法や言語上存在する確からしさ（大語彙音声認識に用いられる言語モデルに示されている）を用いることができる。

【００２１】音声認識結果の単語系列に対して話題抽出を行う際に、認識言語系列候補の第１位だけでなく、上位ｂ位までの候補（$w_{1\text{-}1}$，$w_{1\text{-}2}$，…，$w_{1\text{-}n1}$），（$w_{2\text{-}1}$，$w_{2\text{-}2}$，…，$w_{2\text{-}n2}$）…（$w_{b\text{-}1}$，$w_{b\text{-}2}$，…，$w_{b\text{-}nb}$）を用いて話題抽出を行う。この際、順位の高い程重みが大きくなるようにすることもできる。この場合第１位から第ｂ位までの候補系列は、相互に１単語又は２単語しか違いがない候補系列が多くなる。よってこれら候補系列を、その同一単語を排除して複数単語木構造乃至単語ネットワークあるいは単語ラティスの配列とし、これを用いて第１位～第ｂ位の候補系列について話題抽出をするようにすれば、その複数の候補系列を少ない容量のメモリに格納して処理することができる。

【００２２】

【発明の効果】評価は、ニュース音声の書き起こし文および２万語彙の大語彙連続音声認識システムによる音声認識結果に対してこの発明の評価を行った。書き起こし文に対して３人の被験者が人手で付与した話題を評価対象とした。話題抽出モデルが出力した話題単語のうち上位５単語までを出力結果とした場合の適合率（抽出した話題単語のうち、正解の話題単語の割合）は、３人の被験者の付与した話題に対して７０％以上となった。また、単語誤り率２５％の音声認識結果に対する話題抽出結果の適合率も６５％以上となった。各被験者が付与した話題間の重複は約７０％であるので、この話題抽出結果は利用可能な精度であるといえる。関連度の尤度としてx^2法を用いた方が相互情報量を用いた場合より良い結果が得られた。

【００２３】この発明によれば、大量のテキストデータを用いて非常に多くの文中単語および話題単語間の関連度を学習した話題抽出モデルを用いることにより、テキストおよび誤りを含む大語彙連続音声認識結果から詳細な話題抽出を行うことができるという利点がある。つまり、音声からの話題抽出において、連続音声認識技術を用いることにより、限られた数のキーワードを検出するキーワードスポッティングに基づく方法に比べ、音声中の多くの情報を用いて話題抽出を行うことが可能であり、また、音声の内容を表す単語（話題単語）を複数抽出することにより、音声をいくつかの分野に分類する話題抽出（話題同定・話題認識）に比べ、詳細な話題が抽出できるという利点がある。

【００２４】特に従来のテキストに対する話題抽出では、特定の関係のものを抽出するため、複雑な処理を必要としたが、この発明では比較的簡単に行うことができる。特に連続音声に対する抽出ではその特定部分に対して認識誤りが生じると、致命的であるが、この発明は文全体の各単語に対して関連性をみるため正しく話題を抽出することができる。

【００２５】またこのような正しい抽出ができるのは、大量のテストデータを用いて作成した話題単語と各単語との関連度を記憶した話題抽出モデルを用いるからである。

【図面の簡単な説明】

【図１】この発明のモデル作成方法を示す流れ図。

【図２】Ａはこの発明の話題抽出モデルの例を示す図、Ｂはこの発明の話題抽出方法を示す図である。

【図１】

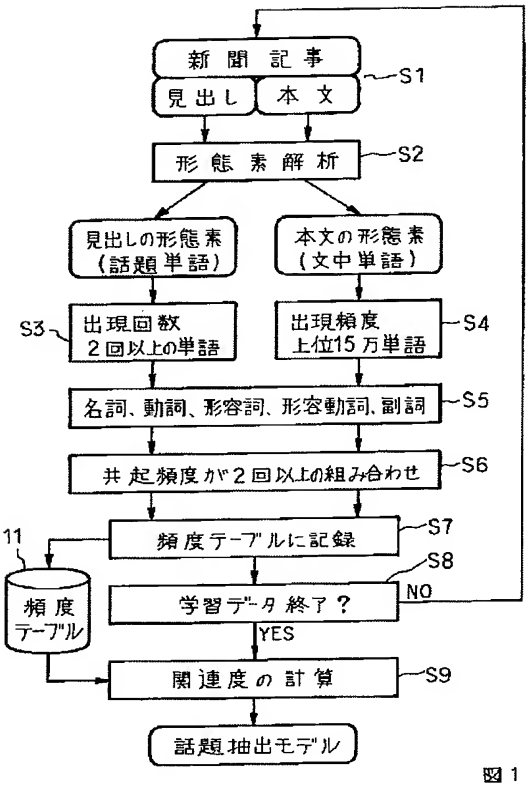

新聞記事
見出し
本文
S1
形態素解析
S2
見出しの形態素（話題単語）
本文の形態素（文中単語）
S3
出現回数2回以上の単語
出現頻度上位15万単語
S4
名詞、動詞、形容詞、形容動詞、副詞
S5
共起頻度が2回以上の組み合わせ
S6
11
頻度テーブルに記録
S7
頻度テーブル
S8
学習データ終了？
NO
YES
関連度の計算
S9
話題抽出モデル

図１

【図２】

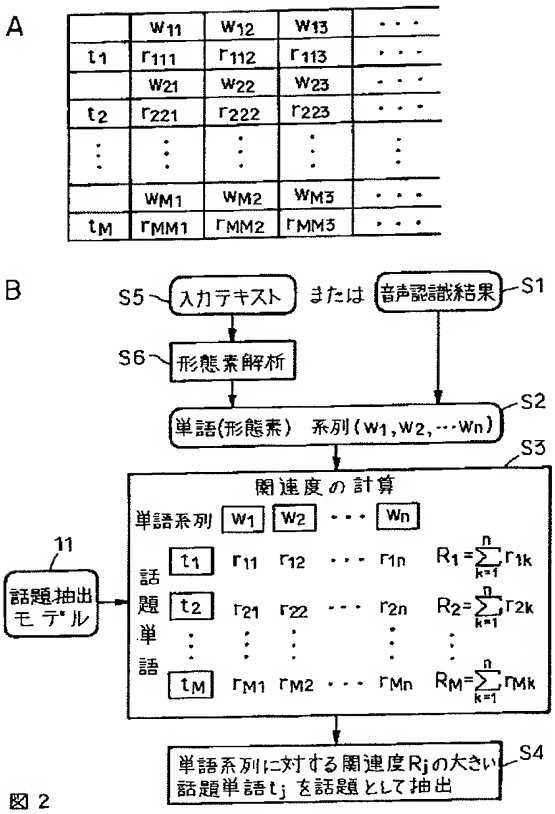

A
B
S5
入力テキスト
または
音声認識結果
S1
S6
形態素解析
単語(形態素) 系列(w1,w2,…wn)
S2
S3
関連度の計算
単語系列
話題単語
11
話題抽出モデル
S4
単語系列に対する関連度Rjの大きい話題単語tjを話題として抽出

図２